ELSEVIER

Discrete Mathematics 138 (1995) 319–326

DISCRETE MATHEMATICS

On path entropy functions for rooted trees☆

A. Meir[a], J.W. Moon[b,*]

[a]*York University, North York, Ont., Canada M3J 1P3*
[b]*Department of Mathematics, University of Alberta, 632 Central Academic Bldg., Edmonton, Alberta, Canada T6G 2G1*

Received 7 July 1993

Abstract

If u is a terminal node of a rooted tree T_n with n terminal nodes, let $h(u) = \sum f(d(v))$ where the sum is over all interior nodes v in the path from the root of T_n to u, $d(v)$ is the out-degree of v, and f is a non-negative cost function. The path entropy function $h(T_n) = \sum h(u)$, where the sum is over the n terminal nodes of T_n, is a measure of the complexity of the hierarchical classification scheme represented by T_n. We show, under suitable assumptions, that the expected value of $h(T_n)$ over all trees T_n in certain families of weighted trees is asymptotic to $Kn^{3/2}$ where the constant K depends on the family and the cost function f.

1. Introduction

The *out-degree* $d(v)$ of a node v in a rooted tree is the number of edges incident with v that lead away from the root. Let T_n denote a rooted tree with n *terminal* nodes (of out-degree zero) and no nodes of out-degree one; we shall refer to such trees as *reduced* trees. Let $f(k)$ denote a non-negative *cost-function* defined for positive integers k. For any terminal node u of a non-trivial reduced tree T_n, let $h(u) = \sum f(d(v))$ where the sum is over the interior nodes v of T_n in the path joining the root of T_n to u. If T_n represents a hierarchical classification scheme with n categories corresponding to the n terminal nodes, then $f(d(v))$ represents the cost of the decision made at the intermediate step corresponding to the interior node v; and the *path entropy function* $h(T_n) = \sum h(u)$, where the sum is over all terminal nodes u of T_n, is a measure of the complexity of T_n.

Green [6] showed that if $f(k) = \log_2 k$, then

$$n \log_2 n \leqslant h(T_n) \leqslant \tfrac{1}{2}(n-1)(n+2)$$

☆ The preparation of this paper was assisted by grants from the Natural Sciences and Engineering Research Council of Canada.

* Corresponding author.

SSDI 0012-365X(94)00214-2

for any reduced tree T_n with $n \geqslant 3$ terminal nodes. (See also [7] or [12] for additional references and material on path functions.) This suggests the problem of considering the expected value of $h(T_n)$ over all trees in specified families $\mathscr{F}$ of reduced trees and for more general cost functions $f(k)$.

In section 2 we describe the families $\mathscr{F}$ of reduced trees we shall be considering — the simply generated families — and we shall determine the asymptotic behaviour of the number of trees T_n in such families. Our main results are in Section 3 where we show that under suitable assumptions the expected value of $h(T_n)$ grows like $Kn^{3/2}$ where the value of the constant K depends on the family $\mathscr{F}$ and the cost function $f(k)$. We conclude with some numerical examples in Section 4.

2. Simply generated families

Let $\mathscr{F}$ denote a family of weighted *plane* trees (or *ordered* trees, as they are sometimes called [8, p. 306]) in which the tree T_n has *weight* $w(T_n)$. We recall (see, e.g., [4, 9, 13]) that such a family is said to be *simply generated* if there exists a sequence of non-negative constants $c_0\ (=1), c_1, c_2, \ldots$ such that

$$w(T_n) = \prod c_i^{D_i(T_n)} \tag{2.1}$$

for all trees T_n in $\mathscr{F}$, where $D_i(T_n)$ denotes the number of nodes of out-degree i in T_n. We shall assume henceforth that $\mathscr{F}$ is some particular simply generated family of trees for which $c_1 = 0$ so that only reduced trees, with no nodes of out-degree one, receive non-zero weights.

Let $y_n = \sum w(T_n)$, where the sum is over the (reduced) trees T_n in $\mathscr{F}$ with n terminal nodes. If we classify the trees T_n in $\mathscr{F}$ according to the out-degrees of their roots and take the weight-factors c_i into account, then it is not difficult to see that the generating function $Y = \sum y_n x^n$ satisfies the relation

$$Y(x) = x + c_2 Y^2(x) + c_3 Y^3(x) + \cdots.$$

Theorem 1. *Suppose* $\Psi(t) = c_2 t^2 + c_3 t^3 + \cdots$ *is an analytic function for* $0 \leqslant |t| < R \leqslant \infty$ *and that*

(i) $c_k \geqslant 0$ for $k \geqslant 2$,

(ii) $\gcd\{k-1: c_k > 0\} = 1$, *and*

(iii) $\Psi'(v) = 1$ *for some* v, *where* $0 < v < R$.

If $Y = Y(x) = \sum_1^\infty y_n x^n$ *is the unique solution of the relation*

$$Y = x + \Psi(Y) \tag{2.2}$$

in the neighbourhood of $x = 0$ *with* $Y(0) = 0$, *then* $\Psi(x)$ *is analytic in the disk* $|x| \leqslant \beta = v - \Psi(v)$ *except at* $x = \beta$. *Moreover,* $Y(\beta) = v$ *and*

$$y_n \sim a\beta^{-n} n^{-3/2} \tag{2.3}$$

as $n \to \infty$, *where* $a = (\beta/2\pi\Psi''(v))^{1/2}$.

Remark. Comtet [1] and Foulds and Robinson [5] have shown that conclusion (2.3) holds when $\Psi(t) = e^t - 1 - t$; see also [11, p. 208]. The more general result can be proved by an argument similar to that used to prove Theorem 3.1 in [9] which describes the behaviour of coefficients in generating functions $P(x)$ satisfying a relation of the form $P(x) = x\Phi(P(x))$; see also [13, p. 32]. In fact, conclusion (2.3) can be deduced from the earlier result if relation (2.2) is rewritten as $Y(x) = x\Phi(Y(x))$ where $\Phi(t) = (1 - \Psi(t)/t)^{-1}$. However, for the sake of completeness, we sketch a proof of (2.3) here that involves a somewhat different approach at the outset.

Proof of Theorem 1. We begin by showing that the mapping $x = f(Y) := Y - \Psi(Y)$ is univalent on the disk $|Y| \leqslant v$. If $|Y_1| \leqslant v$ and $|Y_2| \leqslant v$ but $Y_1 \neq Y_2$, then

$$\frac{f(Y_1) - f(Y_2)}{Y_1 - Y_2} = 1 - \sum_2^\infty c_k(Y_1^{k-1} + Y_1^{k-2}Y_2 + \cdots + Y_2^{k-1}). \tag{2.4}$$

Suppose, first, that $|Y_1| = |Y_2| = v$ and let $e^{i\theta} = Y_2/Y_1$; then

$$\left|\sum_2^\infty c_k(Y_1^{k-1} + Y_1^{k-2}Y_2 + \cdots + Y_2^{k-1})\right| \leqslant \sum_2^\infty c_k v^{k-1} \cdot |1 + e^{i\theta} + \cdots + e^{i(k-1)\theta}|$$

$$< \sum_2^\infty k c_k v^{k-1} = 1,$$

by (i) and (iii), since $|1 + e^{i\theta} + \cdots + e^{i(k-1)\theta}| < k$ if $\theta \neq 0$. Hence $f(Y_1) \neq f(Y_2)$ in this case, in view of (2.4). To dispose of the remaining cases, when $|Y_1| < v$ or $|Y_2| < v$, we may observe that $|Y_1^{k-1} + Y_1^{k-2}Y_2 + \cdots + Y_2^{k-1}| < kv^{k-1}$ and proceed as before; or we may appeal to the result (cf. [2, p. 184]) that if a function is analytic within and on a closed contour and is univalent on the contour, then it is univalent within and on the contour. It follows, therefore, that $x = f(Y)$ is univalent on the disk $|Y| \leqslant v$ in the complex Y-plane. Moreover, f maps the circle $|Y| = v$ onto a simple closed curve Γ in the complex x-plane and f maps the interior of the circle $|Y| = v$ onto the interior of the region enclosed by Γ; in particular, $f(0) = 0$ and $f(v) = \beta$.

Let Y_0 and x_0 be any points such that $x_0 = f(Y_0)$ and, in addition, $|Y_0| = v$ but $Y_0 \neq v$ or, equivalently, $x_0 \in \Gamma$ but $x_0 \neq \beta$. Since conditions (i) and (ii) hold and $v < R$, it follows from Lemma 2 in [10] that $|1 + \Psi(Y_0)/Y_0| < 1 + \Psi(v)/v$ and, hence, that

$$|x_0| = |Y_0 - \Psi(Y_0)| = v|2 - (1 + \Psi(Y_0)/Y_0)|$$

$$> v(2 - (1 + \Psi(v)/v)) = \beta.$$

This implies that the set $D = \{x: |x| = \beta, x \neq \beta\}$ lies in the interior of the region enclosed by Γ, so for every $x \in D$ there is a unique Y such that $|Y| < v$ and $x = f(Y)$.

Furthermore, this inverse function $Y = Y(x)$ is analytic for all $x \in D$ since if $|Y| < v$ then

$$|f'(Y)| = |1 - \Psi'(Y)|$$

$$\geqslant 1 - \Psi'(|Y|) > 1 - \Psi'(v) = 0$$

by (i) and (iii); and $Y(x)$ has a singularity at $x = \beta$ since $f'(v) = 0$.

Now $Y(x) = \sum_1^\infty y_n x^n$ around $x = 0$, so this series converges for $|x| < \beta$. Since the mapping $x = f(Y)$ is continuous it follows that $Y(\beta^-) = v$; the coefficients y_n are non-negative, by (i) and relation (2.2), so it follows from Tauber's theorem that $\sum_1^\infty y_n \beta^n = v$.

The rest of the argument is standard. In the neighbourhood of $Y = v$ we have the expansion

$$\beta - x = (v - Y) - (\Psi(v) - \Psi(Y)) = \tfrac{1}{2}\Psi''(v)\cdot(v - Y)^2 + \cdots;$$

hence in the neighbourhood of $x = \beta$ we have the expansion

$$Y(x) = v - b_1(\beta - x)^{1/2} - b_2(\beta - x)^1 - b_3(\beta - x)^{3/2} - \cdots, \tag{2.5}$$

where, in particular, $b_1 = (2/\Psi''(v))^{1/2}$. Darboux's theorem [3, p. 20] states that if $Y(x)$ has the expansion (2.5) where β is the only singularity of $Y(x)$ on its circle of convergence, then conclusion (2.3) holds where $a = \frac{1}{2}b_1(\beta/\pi)^{1/2} = (\beta/2\pi\Psi''(v))^{1/2}$, as required. (We remark that conclusion (2.3) still holds when $c_1 \neq 0$ and nodes of out-degree one are admitted provided $\Psi(t)$ is defined as $\sum_1^\infty c_i t^i$; note that a necessary condition for condition (iii) to be satisfied then is that $c_1 < 1$.) □

3. Main results

Let

$$h_n = \sum w(T_n)\cdot h(T_n)$$

for $n \geqslant 2$, when the sum is over all trees T_n with n terminal nodes that belong to a particular simply generated family $\mathcal{F}$ of reduced trees, $w(T_n)$ is the weight of T_n defined by (2.1), and $h(T_n)$ is the path entropy function defined in Section 1. Let $F(t)$ and $F'(t)$ denote the formal power series

$$\sum_2^\infty c_k f(k) t^k \quad \text{and} \quad \sum_2^\infty c_k k f(k) t^{k-1},$$

where the c_k's are the weights appearing in the definition of $w(T_n)$ and $f(k)$ is the cost function appearing in the definition of $h(T_n)$. We now derive a relation for the

generating function

$$H(x) = \sum_{2}^{\infty} h_n x^n$$

in terms of $F(t)$ and the generating function $Y(x)$ for the family $\mathscr{F}$.

Theorem 2. $H(x) = xF'(Y(x))\cdot(Y'(x))^2$.

Proof. Let

$$U(x,z) = x + \sum_{n=2}^{\infty} \left(\sum w(T_n) z^{h(T_n)} \right) x^n,$$

where the inner sum is over all trees T_n in $\mathscr{F}$ with n terminal nodes; we note that $U(x,1) = Y(x)$ and $U_z(x,1) = H(x)$. If the tree T_n is formed by joining the root-node to the roots of the subtrees $T^{(1)}, \ldots, T^{(k)}$ for some $k \geqslant 2$, then it follows from the definition of $h(T_n)$ that

$$h(T_n) = nf(k) + \sum_{j=1}^{k} h(T^{(j)}).$$

From this it is not difficult to see that

$$U(x,z) = x + \sum_{2}^{\infty} c_k U^k(xz^{f(k)}, z).$$

If we take the formal partial derivatives of both sides of this relation with respect to z and set $z = 1$, we find that

$$\begin{aligned} H(x) &= \sum k c_k Y^{k-1}(x)\cdot\{xf(k)\,Y'(x) + H(x)\} \\ &= xF'(Y(x))\cdot Y'(x) + \Psi'(Y(x))\cdot H(x). \end{aligned}$$

Hence

$$\begin{aligned} H(x) &= xF'(Y(x))\cdot Y'(x)\cdot(1 - \Psi'(Y(x)))^{-1} \\ &= xF'(Y(x))\cdot(Y'(x))^2 \end{aligned}$$

as required, appealing to relation (2.2) at the last step. □

In what follows we shall use the following straightforward asymptotic result; we let $\mathscr{C}_n\{g(x)\}$ denote the coefficient of x^n in the power series expansion of $g(x)$.

Lemma. *Let* $P(x) = \sum_0^\infty p_n x^n$ *and* $Q(x) = \sum_0^\infty q_n x^n$ *and suppose that*

$$p_{n-1}/p_n \to 1 \quad \text{as } n \to \infty, \tag{3.1}$$

$$p_m/p_n \leqslant H < \infty \quad \text{for } 0 \leqslant m \leqslant n \text{ and } n = 0, 1, \ldots, \tag{3.2}$$

and

$$\sum_{0}^{\infty} |q_n| = A < \infty . \tag{3.3}$$

Then

$$\mathscr{C}_n\{Q(x)P(x)\} \sim Q(1)p_n \quad \text{as } n \to \infty . \tag{3.4}$$

Proof. We may assume without loss of generality that $p_n > 0$ for $n \geqslant 0$. For any given $\varepsilon > 0$ we may choose $M = M_\varepsilon$ so that $\sum_{M+1}^{\infty} |q_n| < \varepsilon$, by (3.3); and then we may choose $N = N_\varepsilon$ so that $|p_{n-k}/p_n - 1| < \varepsilon$ if $0 \leqslant k \leqslant M$ and $n \geqslant N$, by (3.1). If $n \geqslant N$, then

$$\begin{aligned} &|p_n^{-1} \cdot \mathscr{C}_n\{Q(x)P(x)\} - Q(1)| \\ &\quad \leqslant \left| \sum_{0}^{M} q_k(p_{n-k}/p_n - 1) \right| + \left| \sum_{M+1}^{n} q_k p_{n-k}/p_n \right| + \left| \sum_{M+1}^{\infty} q_k \right| \\ &\quad \leqslant \varepsilon(A + H + 1), \end{aligned}$$

where we have used (3.2) at the last step. The required result now follows. □

We now determine the asymptotic behaviour of the expected value $\mu(n) = h_n/y_n$ of the path entropy function $h(T_n)$ over the y_n trees in $\mathscr{F}$. We assume the assumptions of Theorem 1 hold so that $y_n \sim a\beta^{-n}n^{-3/2}$ where $\beta = v - \Psi(v)$, $\Psi'(v) = 1$, and $Y(\beta) = v$.

Theorem 3. *If the series* $\sum_2^{\infty} c_k k f(k) v^{k-1}$ *converges and has sum* S, *then*

$$\mu(n) \sim Kn^{3/2} \tag{3.5}$$

as $n \to \infty$, *where* $K = (\pi/2\beta\Psi''(v))^{1/2} \cdot S$.

Remark. The series defining S will converge if $f(k) = O((1 + \delta)^k)$ for some fixed δ such that $0 < \delta < R/v - 1$ or, in particular, if $f(k) = O(k^N)$ for some fixed N. If the series defining S diverges, then $\mu(n)$ will grow more rapidly than $n^{3/2}$, but we shall not pursue this case further here.

Proof of Theorem 3. We recall that it follows from relation (2.3) and Lemma 3.1(ii) in [9] that

$$\mathscr{C}_n\{(Y'(x))^2\} \sim \pi(a/\beta)^2 \cdot \beta^{-n}.$$

It is not difficult to verify that the functions $P(x) = (Y'(x\beta))^2$ and $Q(x) = x\beta \sum_2^{\infty} c_k k f(k) Y^{k-1}(x\beta)$ satisfy the conditions of the lemma. Thus it follows from Theorem 2 and the lemma that

$$h_n \cdot \beta^n \sim S \cdot \pi a^2/\beta.$$

Hence, $\mu(n) = h_n/y_n \sim Kn^{3/2}$ where $K = S\pi a/\beta = (\pi/2\beta\Psi''(v))^{1/2} \cdot S$, as required. □

Notice that if the weight function $f(k)$ equals one for all $k \geqslant 2$, then $F(t) = \Psi(t)$ and $F'(v) = \Psi'(v) = 1$. In this case conclusion (3.5) implies that the expected value of the sum of the distances from the n terminal nodes of T_n to the root of T_n is asymptotic to $(\pi/2\beta\Psi''(v))^{1/2} n^{3/2}$ as $n \to \infty$. (In particular, if $F(t) = \Psi(t) = t^2$ and $\mathscr{F}$ is the family of binary trees, then $\mu(n) = 4^{n-1}/y_n - n$; see [8, p. 590].) An analogous result on the sum of the distances between the root and *all* nodes in (not necessarily reduced) simply generated trees was given in [9, Theorem 4.5].

4. Some numerical results

Let $\mathscr{F}$ denote the family of reduced 2–3 trees in which the out-degree of every interior node is two or three and $c_2 = c_3 = 1$. Then $\Psi(t) = t^2 + t^3$, so $v = 1/3$ and $\beta = 5/27$, and it follows from Theorem 1 that

$$y_n \sim a(5.4)^n \cdot n^{-3/2},$$

where $a = (5/(8 \cdot 27\pi))^{1/2} = 0.0858\cdots$. (We remark that this family was considered in [4] in connection with another problem; but there these trees were enumerated by the total number of nodes instead of the number of terminal nodes.)

Next, let $\mathscr{F}$ denote the family of reduced plane trees for which $\Psi(t) = t^2(1-t)^{-1}$. In this case $v = 1 - \sqrt{2}/2$, $\beta = 3 - 2^{3/2}$, and

$$y_n \sim a(5.828\cdots)^n \cdot n^{-3/2},$$

where $a = ((3 - 2^{3/2})/2^{7/2}\pi)^{1/2} = 0.0694\cdots$.

Finally, let $\mathscr{F}$ denote the family of reduced labelled trees for which $\Psi(t) = e^t - 1 - t$. (This is the family enumerated in [1, 5], as we mentioned before.) Then $v = \ln 2$, $\beta = \ln(4/e)$, and

$$y_n \sim a(2.588\cdots)^n \cdot n^{-3/2},$$

where $a = (\ln(4/e)/4\pi)^{1/2} = 0.1753\cdots$. In Table 1 we illustrate the conclusion of Theorem 3 on these three families for five particular cost functions; the limiting values have been truncated after three decimal places.

Table 1
Some values of $K = \lim_{n\to\infty} \mu(n) \cdot n^{-3/2}$

Family	Cost function				
	$f(k) = 1$	$f(k) = \log_2 k$	$f(k) = k$	$f(k) = k^2$	$f(k) = 2^k$
Reduced 2–3 trees	1.456	1.740	3.397	8.251	7.766
Reduced plane trees	1.272	1.695	3.379	10.214	12.285
Reduced labelled trees	1.425	1.726	3.402	8.726	8.555

Note added in proof. F. Göbel and C. Hoede have considered the problem of determining the structure of trees T_n that minimize $h(T_n)$ for certain cost functions $f(k)$ in: "On an Optimality Property of Ternary Trees", Information and Control 42 (1979) 10–26.

References

[1] L. Comtet, Sur le quatrième problème et les nombres de Schröder, C.R. Acad. Sci. Paris 271 (1970) 913–916.
[2] E.T. Copson, An Introduction to the Theory of Functions of a Complex Variable (Oxford, 1935).
[3] G. Darboux, Mémoire sur l'approximation des fonctions de très grands nombres; et sur une classe étendu de développements en série, J. Math. Pures Appl. 4 (1978) 5–56.
[4] P. Flajolet and A. Odlyzko, The average height of binary trees and other simple trees, J. Comput. System Sci. 25 (1982) 171–213.
[5] L.R. Foulds and R.W. Robinson, Enumeration of phylogenetic trees without points of degree two, Ars Combin 17A (1984) 169–183.
[6] C.D. Green, A path entropy function for rooted trees, J. ACM 20 (1973) 378–384.
[7] C.D. Green and F. Suraweera, A. path entropy function for rooted acyclic digraphs, J. Univ. Kuwait (Sci.) 12 (1985) 15–20.
[8] D.E. Knuth, The Art of Computer Programming, Vol. I (Addison-Wesley, Reading, MA, 1973).
[9] A. Meir and J.W. Moon, On the altitude of nodes in random trees, Canad. J. Math. 30 (1978) 997–1015.
[10] A. Meir and J.W. Moon, The asymptotic behaviour of coefficients of powers of certain generating functions, European J. Combin 11 (1990) 581–587.
[11] J.W. Moon, Some enumerative results on series–parallel networks, Ann. Discrete Math. 33 (1987) 119–226.
[12] J. Nievergelt and C.K. Wong, Upper bounds for the total path length of binary trees, J. ACM 20 (1973) 1–6.
[13] J.-M. Steyaert and P. Flajolet, Patterns and pattern-matching in trees: an analysis, Inform. and Control 58 (1983) 19–58.